CALIFORNIA STATE SCIENCE FAIR
2009 PROJECT SUMMARY

Name(s)

Akshay J. Maheshwari

Project Number

S1612

Project Title

Zergling: An Optimizing Expert System for High Speed Detection of Chimeras Formed during PCR Amplification of 16S rRNA

Abstract

Objectives/Goals

16s rRNA is a region of highly conserved RNA found in ribosomes that is extremely important in the identification of organisms. Through the process of PCR amplification, RNA from this region can be sequenced and analyzed. Although PCR is a common and frequently used technique, the process is not perfect - the genes from different organisms are often accidentally spliced together to make Chimeras. Chimeras are artificial artifacts of PCR and the RNA does not exist in any extant organism. The current most common way to get rid of the Chimeras from databases is to have scientists manually check each of the hundreds of thousands of sequences for chimeric qualities. The goal of this project was to create a program, Zergling, which could efficiently and accurately identify and eliminate these chimeras from the data set.

Methods/Materials

Zergling solves the problem of Chimeras through a linear algorithm rather than the quadratic algorithms advocated by other chimera checkers. It does so by utilizing a reference database of 2000 known sequences and a percent composition dual-bagging algorithm. The base Reference Database was created by manually going through RNA databases and collecting mean representations of different taxa; as new sequences are processed, the database can learn and grow. The Percent composition and Dual-bagging algorithm compute 5 numbers from each 12,000 index input sequence enabling efficient and accurate comparison against each reference database sequence. This results in an O(n) speed with an insignificant constant k, unlike programs such as Mallard that run at O(n2) speed and with extremely high constants.

Results

The speed and accuracy benchmarks of Zergling were compared against those of Mallard and the manual checking by phylogeneticists. Zergling consistently processed sequences more than 100x faster than Mallard and more than 500x faster than manual checking. Zergling was highly accurate and found within 95% of the chimeras found by manual checking while Mallard averaged under 75%.

Conclusions/Discussion

Zergling greatly increased the speed and accuracy of chimera identification. Because of this, it has the capability to process PCR sequences in real time as well as screen millions of sequences from past databases in a batch process. Zergling can learn from new sequences and will be used to validate past and future databases. It will soon be released for use by scientists worldwide.

Summary Statement

This project identifies erroneously spliced genes formed during PCR. A novel learning algorithm is utilized which makes batch processing of RNA feasible in real time.

Help Received

My Mother and Father helped design poster; Bioinformatics Research Associate Elisabeth Bik assisted in the validation of my project